Examining High and Low Value-Added Mathematics Instruction:

Can Expert Observers Tell the Difference?

Heather C. Hill

David Blazar

Harvard Graduate School of Education

Andrea Humez

Boston College

Erica Litke

Harvard Graduate School of Education

Mary Beisiegel

Oregon State University

Johanna Barmore

Mark Chin

Harvard Graduate School of Education

Douglas Corey

Brigham Young University

Sara Roesler

Lexington Public Schools

Lucas Salzman

University of Pennsylvania

David Braslow

Samantha Rabinowicz

Harvard Graduate School of Education

Paper presented at Association for Public Policy Analysis and Management Annual Conference, Washington DC, November 7, 2013. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C090023 to the President and Fellows of Harvard College to support the National Center for Teacher Effectiveness. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

The question of how to measure effective teachers and teaching has long been of interest to policymakers and school leaders (Fenstermacher & Richardson, 2005; Peterson, 2000; Stodolsky, 1988). While recent policy initiatives have focused on the use of value-added measures (VAM) to assess teacher quality, there is a much longer tradition of using observations of practice to make such determinations (Brophy & Good, 1986; Cooley & Leinhardt, 1980). However, empirical evidence suggests these two indicators often identify different sets of teachers as effective. For example, the Measures of Effective Teaching project (Kane & Staiger, 2012) finds low correlations between teachers' VAM scores and their quality of instruction as measured by observational metrics. Studies with the explicit intent of identifying differences in instruction between teachers with high and low VAM scores (Grossman et al., in press; Stronge, Ward, & Grant, 2011) also have generally failed to uncover substantial differences across classrooms.

One reason for this disconnect may be the nature of the educational production function. Although scholars have spent the better part of the last three decades studying the relationship between teachers, teaching, and student outcomes (e.g., Brophy & Good, 1986; Wayne & Youngs, 2003), there is little agreement around which classroom- or teacher-level factors lead to student success. Instead, scholars have identified dozens of variables – classroom management, classroom climate, teacher knowledge, teachers' use of assessments, specific instructional practices, teachers' use of specific kinds of curriculum materials – each with low to moderate correlations with student outcomes. Because each line of inquiry has occurred independently, often in different eras, conclusions regarding the relative efficacy of each factor have been difficult to ascertain.

Recently, decreases in the relative costs of obtaining both video records of instruction and value-added scores have made investigations into the correspondence between these two measures more practical. In this study, we take advantage of a dataset containing both videotaped lessons and value-added scores to mount an exploratory study of the instruction of teachers with high- and low-value-added rankings. Specifically, we seek to answer two questions: First, what is the degree of convergence between observers' impressions of mathematics instruction and teachers' mathematics value-added scores? Second, are there a set of instructional practices that consistently characterize high but not low-value-added ranked teachers' classrooms, and vice versa?

To answer these questions, we use data generated by fourth- and fifth-grade math teachers and their students in three large public school districts. After ranking teachers within districts based on a standard value-added model, we identified three teachers in each of the lowest, middle, and top quintiles for further analysis (*n*=27 across three districts). Observers blind to those VAM rankings then viewed a minimum of five videotaped lessons from each teacher, individually providing lesson-level ratings on a broad set of mathematical and pedagogical instructional features and qualitatively describing strengths and weaknesses in these areas. Next, observers convened to predict each teacher's value-added score quintile, to rank order teachers within district, and to record the instructional features prevalent in each classroom.

Our analyses indicate only modest convergence between raters' and VA models' rank order estimates. In fact, in roughly half of cases, the observation group incorrectly guessed teachers' VAM quintiles. Discrepancies were severe in one of the three districts, where observers identified three-quarters of teachers incorrectly. An exploratory analysis of instructional features suggests that several were associated with value-added outcomes, including the efficiency with which the teacher managed the classroom, the density of the mathematics, the clarity of the mathematics, and the overall mathematical quality of instruction (all $p < 0.10$, $n=27$). A comparison of high- and low-scoring teachers based on observer-written memos, lesson transcripts, and other artifacts suggested, however, that it was not possible to disentangle the relative importance of each of these instructional features – or other possible factors not captured through observation of lessons – to teachers' ability to raise student test scores.

Below, we provide the motivation for this analysis. Next, we describe our cross-district results. Specifically, we describe the overall quality and variability of instruction within each, the process by which raters came to agreement about teachers' predicted value-added rankings, and the degree of convergence between predictions and actual rankings. Based on these results, we explore instructional features that characterize high- or low-ranked teachers and, in cases where there is misalignment, hypothesize possible explanations for these discrepancies. Finally, we discuss the implications of these findings for policy and practice.

Literature review

Despite several decades of concerted inquiry, scholars have only partially explained the "production function" that converts classroom teaching into student outcomes. For instance, in studies that associate observations of practice with student learning outcomes, the correlation rarely exceeds 0.3 (Bell, Gitomer, Hamre, Pianta, & Qi, 2012; Hill, Kapitula, & Umland, 2011; Jacob & Lefgren, 2006; Milanowski, 2004; for an exception see Schacter & Thum, 2004). Studies that use surveys to describe teacher or teaching characteristics – for instance, teachers' mathematical preparation, experience, certification, attitudes, and self-reported instruction – explain on average 8-10% of variation in student outcomes (Boonen, Van Damme, & Onghena, 2013; Palardy & Rumberger, 2008). In both cases, the majority of teacher-level variation in student outcomes is left unexplained.

Explaining Weak Relationships Between Teacher Characteristics and Value-Added

One reason for these modest correlations may be value-added scores themselves, which are relatively noisy measures of teacher effectiveness. For instance, a series of studies have estimated the adjacent-year correlations between teacher VAM scores to be in the 0.2 to 0.5 range (Aaronson et al., 2007; Koedel & Betts, 2007; McCaffrey et al., 2009). This finding has led economists to recommend using models that calculate teacher scores using multiple years of data (Goldhaber & Hansen, 2012; Koedel & Betts, 2010). Goldhaber and Hansen (2012), for instance, find that adjacent-year correlations in North Carolina panel data average 0.55; however, when adjacent three-year averages are correlated, that figure rises to 0.65. Despite these improvements in stability, however, it remains clear that at least some portion of value-added scores consist of measurement error – potentially owing to error inherent in the way student tests are scored and used (Koedel, Leatherman, & Parsons, 2012), or to disturbances at the teacher level (e.g., the proverbial barking dog or single miscreant student, see Kane, Staiger, Grissmer, & Ladd, 2002). Measurement error would tend to attenuate relationships between teaching and student outcomes. Although in one recent study (Kane & Staiger, 2012), even correcting for such measurement error returned some of the lowest observed estimated correlations between VAM scores and scores from observational metrics.

Another reason for these modest correlations may be noise in the observational instruments. Hill, Charalambous, and Kraft (2012) estimate, in a small g-study with a fully crossed design, reliabilities for the *Mathematical Quality of Instruction* (MQI) instrument at roughly 0.80 for two raters each scoring four lessons. The MET study (Kane & Staiger, 2012) reports much lower estimated reliabilities, ranging from, for four lessons, 0.20 to 0.68 for a variety of instruments, including the MQI. Bell and colleagues (2012) do not estimate reliabilities for different configurations of raters and teachers, but report less variation at the teacher level than Hill and colleagues (2012). If the reliability of teacher observations is in the neighborhood, on average, of 0.6, then correlations between observational and VAM scores would be significantly attenuated by measurement error.

A third reason may be that research on teaching generally occurs in silos. Researchers tend to examine the relationship between only one or two teaching traits and student outcomes. Yet, the production function may be multidimensional and complex. For instance, reviews of the effective teaching literature (Brophy, 1999; Stronge et al., 2011) suggest that effective instruction is the sum – or possibly the interaction of – many facets of pedagogical practice. Thus, measuring only a limited number of inputs will result in weak explanatory power in any given study and a literature that features, across studies, many small effects.

A fourth explanation may lie in what cannot be measured or is difficult to measure observationally. For instance, test preparation activity tends to occur at specific times of the year – thus eluding many observation schedules – and may in fact be difficult to identify for all but the most familiar with the formats and content of the test. In addition, effective teaching practices may be contextually bound – e.g., Practice A works well in situation X with Y population of children – making effective teaching more about the application of particular practices in the right moment than the use of any specific practice over another. As conventional observation instruments and video observation protocols are limited to describing the occurrence, rather than their appropriateness, of practice, they may be unable to measure what matters.

Instructional Characteristics of Effective Teachers

Despite historical inability to explain substantial variability in student achievement through teacher/teaching characteristics, recent use of observational rubrics in large-scale datasets has allowed researchers to understand and compare the effects of a broader range of instructional features. Results from this work provide a clearer picture, consistently pointing to stronger relationships to student achievement for classroom climate and management than other instructional features. Stronge and colleagues (2011) find that variables focused on classroom climate and management best differentiate between teachers with high and low value-added scores. Bell and colleagues (2012) find that the classroom organization scale of the CLASS best predicts student gains in high school algebra classrooms. Tyler, Taylor, Kane and Wooten (2010) find that, in Cincinnati, having relatively better scores on the classroom management dimension than the instruction dimension of Framework for Teaching predicts student outcomes in mathematics and reading; the contrast between inquiry-oriented practices and routine instruction is only significant for reading. In ELA, Grossman and colleagues (2010) find explicit strategy instruction and student engagement to differentiate between teachers with high and low value-added scores. These findings are striking given that while most of these instruments contain items describing inquiry-oriented instruction, none appear significant. At the same time, the bulk of these studies have occurred with formal coding schemes in place of, rather than via exploratory analyses, and thus might not capture important differences between classrooms of teachers with higher and lower value-added score.

These emergent results are particularly informative at a time when two initiatives seek to improve teaching and learning. The first is the Common Core State Standards, which in mathematics calls for more discipline-grounded thinking and reasoning on the part of students. This continues a tradition of inquiry-oriented reforms in mathematics (NCTM, 2000), yet the findings above suggest that at least as measured on current instruments and against current tests, such instruction rarely fares well. The second reform focuses on teacher evaluation; in most states, new, more discriminating teacher scores will be based on a combination of new observational rubrics and value-added metrics. Although the combination of two weakly related measures may complicate evaluation efforts (Chester, 2003; Martinez, 2012), it is thought that value-added and principal evaluations converge in the upper and lower tails (Jacob & Lefgren, 2008), important given that this is where most rewards and sanctions are targeted.

Research Questions

This study aims to illuminate several issues that emerge from prior work. First, it will estimate the degree of convergence between raters' observations of videotaped instruction and teacher value-added scores. Doing so will shed light on the contention that observations and value-added scores converge for teachers in the top and bottom quintile of the value-added distribution. We explore this issue without a formal coding scheme, instead developing hypotheses about differences among teachers' classrooms from an exploratory round of coding, then using new items generated from these hypotheses as well as observer notes to rank teachers by the quality of their instruction. Both the exploratory and analytic phase of the study may help generate hypotheses regarding the features that appear in the instruction of teachers with high- value-added scores, but not in those with low-value-added scores. This process may also provide insight into whether classroom climate and management are the major factors predicting value-added outcomes or whether there are other, disciplinary or inquiry-based features that also help to do so. Using an open coding scheme allows us to examine a variety of instructional characteristics that may be related to student achievement. Answering these questions together may provide insight into the nature of the educational production function – to what extent student outcomes on standardized tests can be explained, and whether any one specific facet or multiple facets together provide the most explanatory power.

Methods

Data

For this study, we draw primarily on administrative data and video records of instruction from teachers in three school districts on the east coast of the United States (henceforth "B", "G", and "R"). Administrative records include teacher-student links, demographic information, and end-of-year mathematics test scores in the 2009-10 through 2011-12 school years for all fourth- and fifth-grade students in these districts.[1] Video data come from a subsample of these teachers who agreed to participate in a multi-year project aimed at understanding measures of effective mathematics teaching.[2] We recruited schools based on referrals from district leaders and school size, requiring that schools have at least two teachers at each of the sampled grades. Across all districts in the study, 55% of teachers in these schools agreed to participate. These teachers were allowed to select the dates for video recording in advance; we only required that they select a typical lesson and exclude days on which students were taking a test or preparing for a state standardized test. Lessons were approximately one hour in length on average, with individual lessons ranging from 45 to 80 minutes. As part of the project, these teachers also took a survey that included a test of their Mathematical Knowledge for Teaching (MKT). Table 1 provides descriptive information about our analytic (n=27) sample as compared to the broader study sample (n=247) and the total sample of fourth and fifth graders in each of our three school districts. Average student characteristics in the analytic and study sample are very similar to those in the districts as a whole.

Sampling

In order to select the subsample of teachers whose lessons we observed in this study, we began by ranking all teachers within each district using a common value-added model:

$$A_{ijcgst} = \zeta\big(f(A_{ijcgst-1})\big) + \pi X_{it} + \varphi C_{cjt} + \alpha S_{st} + \sigma_{gt} + \mu_{cjt} + \gamma_t + \delta_j + \varepsilon_{icjgst}$$

Here, student *i*'s end-of-year test score with teacher *j* in class *c*, grade *g*, school *s*, and year *t*, is modeled as a cubic function of students' prior-year test scores in both math and reading, $f(A_{ijcgst-1})$; a vector of student characteristics, X_{it}, including gender, race, free- or reduced-price lunch eligibility, special education status, and limited English proficiency status; and class

[1] One district also has a fourth year of administrative data from 2008-09, which we include in our value-added model below.

[2] This project also includes a fourth district that is not included in the present study, as the sample of teachers is too small to ensure sufficient numbers of teachers in each value-added quintile.

(C_{cjt}) and school (S_{st}) characteristics aggregated up from the student level.[3] A set of grade-by-year fixed effects, σ_{gt}, controls for the fact that tests differ across grades and school years. To account for the nested structure of the data, we included random effects for classes, μ_{cjt}, school years, γ_t, and teachers, δ_j. Teachers' value-added scores were constructed from δ_j, the best linear unbiased predictor of the teacher-level residual. Ideally, this estimate is a measure of the residual in students' test scores that cannot be explained by prior achievement, background characteristics, class, or school, and therefore is attributed to the teacher. We run models separately for each district using all years of available test-score data to increase the precision of our value-added estimates (Schochet & Chiang, 2013; Koedel & Betts 2011; Goldhaber & Hansen, 2012). For the same reason, we also limit our sample of interest to those teachers with three years of test scores.

Next, within each district we randomly selected three teachers from each of the top, middle, and bottom quintiles of value-added scores, for a final analytic sample of 27 teachers.[4] While prior work has focused only on observing teachers in the tails of the distribution (e.g., Stronge, et al., 2011), we include teachers at all points in order to mirror policy-relevant scenarios that often seek to differentiate teachers between two adjacent rankings, i.e., "low" and "mid", or "high" and "mid" (e.g., Dee & Wyckoff, 2013). We specifically excluded teachers from the second and fourth quintiles in order to increase the chances raters would observe sharp differences among the three sets of teachers. To be selected into the analytic sample, teachers were required to have two years of study data and have at least five videos collected over two years.

Data Analysis

In order to answer our research questions, trained raters of mathematics instruction who were blind to teachers' value-added rankings viewed each teacher's lessons and qualitatively analyzed instruction. Though many of these raters were trained on the *Mathematical Quality of Instruction* instrument, we chose instead to conduct an exploratory round of coding (described below) to generate hypotheses and a set of codes specific to this paper. To score the final analytic sample of 27 teachers, the set of twelve raters were split into three balanced groups that took into account their experience observing instruction. Then, intact groups were assigned randomly to a district. Raters watched all available videotaped lessons for each of the nine teachers in their assigned district blind to the teachers' value-added quintile. The order of teachers, the order of lessons, and the raters assigned to each lesson (generally, three observers per lesson) all were generated randomly.

[3] Because we exclude classes with more than 50 percent of students classified as special education, we do not include this as a class-level covariate. We also exclude classes with fewer than five students and those with more than 50 percent missing prior-year math score, as well as students who were retained in grade.

[4] One district had just three teachers with video data at two of three levels (i.e., low, mid, and high). Other district-level combinations had between five and 22 teachers from which we draw our random sample of three teachers for analysis.

To record results of the viewing process, raters scored teachers on the set of items developed during the earlier round of exploratory analysis. During that exploratory analysis, raters watched videos of teachers ranked either high or low on value-added scores from four districts: the three districts included in this study and one additional district with incomplete district data.[5] Watching video of these teachers, raters noted instructional features that were prevalent across classrooms and that raters felt might explain teachers' scores, then created a set of codes to describe those features. Some codes are specific to the mathematics in the lesson (e.g., "Classroom is Characterized by Mathematical Inquiry", "Mathematics of the Lesson is Clear and not Distorted"), while others focus on more general teaching practices such as student-teacher interactions (e.g., "Teacher Uses Student Ideas") and time-on-task (e.g., "Students are Engaged", "Lesson Time is Used Efficiently") (see Table 2 for full list of items). Scores for each item range from Low (1) to High (5). We designed these items such that the mid-point score (3) represents typical/common practice. For instance, on the code "classroom is characterized by mathematical inquiry," a score of mid denotes a classroom in which students occasionally offer a mathematical explanation or reason and engage in mostly low or moderately low-level cognitive tasks. For "teacher remediates student errors," a score of mid indicates a teacher who consistently corrects student errors but does so in a pro forma way, for instance by restating the question until a student answers it correctly or re-outlining a problematic procedure. By designing the items in this way, we hoped scores would vary across lessons and classrooms.

In the analysis phase of the project, raters scored the lessons from their assigned districts' teachers on these codes, as well as on a summary code from the Mathematical Quality of instruction (MQI) instrument: raters' estimate of the overall MQI for the particular lesson. Overall MQI was rated on a 5-point scale, with a score of 1 or 2 indicating a lesson that is at least somewhat problematic because of teacher mathematical errors, time spent off-task, unclear lesson goals or a lack of correspondence between lesson goals and student activities. MQI scores of 3 indicate a lesson that is neither positive nor negative. Higher scores indicate more mathematically rigorous, student-centered and inquiry-oriented instruction.

In order to capture other elements of instruction, raters also recorded qualitative descriptions of each lesson including the lesson topic, a brief narrative describing the lesson, mathematical issues that emerged, mathematical strong points, and other general thoughts and themes. Next, raters met in their district groups to discuss each teacher individually. The groups reviewed the scores on the exploratory items and discussed points of disagreement; however, they did not reconcile these scores. Instead, each group used raters' individual notes to compile a list of instructional features characterizing each teacher. Group members then came to consensus on a prediction of the teacher's value-added ranking. After following this process for all nine teachers, each district group met a final time to predict the rank order of all of their district's teachers from highest (1) to lowest (9). This allowed groups to re-consider initial rankings in context of lessons and teachers viewed later in the process. It also ensured that the

[5] Because teachers were selected randomly for both the exploratory and analysis phases, there was little overlap in the sample of teachers. One rater saw one teacher in both the exploratory and final rounds of lesson coding. Otherwise, the teachers and rater-teacher combinations were different between rounds.

final rankings included three teachers at each of the three levels (low, mid, and high), which was not necessarily the case while raters were simply watching and characterizing instruction. After finalizing and submitting their rankings to a member of the project staff, each district group received a list of the actual value-added rankings of each teacher in their district, as well as additional information about each teacher (e.g., their MKT score) and classroom composition (e.g., average achievement at the beginning of the year, percent of low-income and limited English proficiency students). In instances where the guesses and actual rankings differed, groups discussed and considered possible explanations.

Finally, the three district groups came together for a series of follow-up meetings to discuss trends both within and across districts and to identify any additional key instructional features groups felt differentiated high from low VAM teachers. Raters examined notes, transcripts and, where necessary, short clips from high and low value-added teachers to determine whether they had missed any features of instruction distinguishing teachers from these two groups. By using two rounds of hypothesis generation – first during the exploratory analysis with a preliminary dataset, then after the identities of high- and low-VAM teachers had been revealed – we hoped to maximize the possibility we would detect characteristics specific to the high- and low-VAM teachers' classrooms. By ranking teachers prior to revealing their VAM scores, we hoped to uncover the extent to which observer ratings would converge with VAM.

Results

Sample of Teachers and Lessons

We begin by describing characteristics of the sampled teachers and lessons. Across all 27 teachers, the average tenure in classrooms was 11.7 years (range 5-26 years), 74% of teachers were female, and 30% were non-white. Students in these classrooms were 74% non-white, 20% ELL, and 67% FRPL. These statistics match the broader student sample from which this subsample was constructed (74% non-white, 20% ELL, 61% FRPL). The average MQI score for sample teachers was 3.25, which is significantly higher than the scores given by raters in the larger sample (2.96, *n*=255).

These three districts had very distinct profiles in terms of instruction. In District B, raters identified a high degree of variability in the quality and depth of mathematics offered to students. This variability was reflected in teachers' Overall MQI scores, which range from 2.48 to 4.11 (on scale of 1 to 5) with a standard deviation of 0.55 (see Table 3). One notable element of instruction was use of inquiry-based curriculum materials that often focused lessons on mathematical sense-making (e.g., connecting multiple strategies for multiplying whole numbers) and relationships between concepts (e.g., correspondences between decimals, fractions, and percents). At the same time, different teachers implemented these lessons with varying degrees of quality and depth – some explaining, for example, why distinct multiplication strategies work and the connections between them, and others doing so in a procedural manner with little attention to the meaning behind the procedures. These differences are evident in scores on some of the exploratory items, with two teachers scoring consistently high (at 4 or above) on use of student ideas and mathematical inquiry, and two teachers scoring low (at 2.5 or below) on these same codes. Although raters noted imprecisions in many of the teachers' instruction, there were only a few instances in which teachers made content errors; these issues tended not to detract from the mathematical point of the lesson.

In District G, instruction was of moderate quality and fairly uniform across teachers. Teachers had overall MQI scores of 2.6 to 3.5, with two-thirds of teachers' scores falling between 2.5 and 3 (see Table 3). Lessons differed in mainly superficial ways. For example, some teachers presented new concepts, which students then practiced in groups or individually, while other teachers used centers and only provided direct instruction in small groups. However, most lessons were similar in that material was directly presented to students rather than developed through inquiry-oriented methods, there were few behavior management issues or serious mathematical errors, and in the fact that only sporadic attention was paid to mathematical meaning-making.

In District R, observers noted that instruction was of poor to moderate quality – three teachers had overall MQI scores of below 2 on the 5-point scale, suggesting moderately to severely problematic mathematics lessons. All teachers' overall MQI scores were below 3.5 (where a score of 3 denotes an average lesson). No individual lesson in the district scored a 5 from all raters, and only 9% scored a 4 or higher. Descriptive notes suggest that lessons in this district featured little student talk and, even when taught by teachers ranked higher by observers, occasional mathematical errors. Sense-making occurred sporadically, and raters did not feel lessons were well-designed, in the sense that even in the best of lessons, teachers seldom brought coherence and completeness to the mathematical topics under study.

Convergence Between Observers' Impression of Instruction and Value-Added Scores

Next, we describe the degree of convergence between observers' impressions of instruction and value-added scores. To frame this discussion, first we discuss estimates of precision and reliability for each of these measures, which could affect results. With regard to value-added scores, we present value-added point estimates and 95% confidence intervals for the nine teachers in our sample (see Figure 1). In all districts, the plausible range for each teacher's point estimates is considerable – in the neighborhood of 0.5-0.6 standard deviations of student achievement. In two of the three districts, some clustering of teachers into the high, medium, and low quintiles – as sampled by design – can be observed. However, in most districts only two or three teachers have non-overlapping confidence intervals, suggesting that even though we used multiple cohorts of student data, these scores are noisy. Decompositions of variance suggests that, controlling for prior student achievement and student and classroom-level demographics, 16% (District B), 9% (District G), and 15% (District R), of the variation in student scores lies at the teacher level.

With regard to observations, we do not have sufficient data or crossing for a formal generalizability study; instead, we provide estimates of variability between raters and variability between lessons within teachers. In Table 4, we show overall and by-district within-one agreement rates for each of the 12 quantitative codes used in this study. These off-by-one agreement rates were strong, suggesting that while raters did not rate lessons uniformly, raters were rarely off by more than one score-point on the 1-5 scale used to score the holistic codes. In Table 5, we show a variance decomposition of these holistic lesson codes by district and teacher. Results suggest very little district-level variability in teachers' scores, but modest amounts of teacher-level variability in comparison to the residual, which contains both lesson-within-teacher variability and measurement error.

Raters predicted value-added ranking less well than expected based on findings in the literature. In roughly half of cases, the observation group incorrectly guessed the quintile within which teachers' VAM score fell (see Table 6). Discrepancies were severe in District B, where observers identified three-quarters of teachers incorrectly and swapped categories (high to low, or vice versa) for three teachers. In Figure 2, we demonstrate the discrepancies between predicted and actual rankings across all three districts. In District G, raters incorrectly predicted value-added rankings for four teachers – two initially ranked high and two initially ranked low – swapping categories in all four instances. In District R, raters were able to ascertain a nearly correct order. Only two teachers were placed in the incorrect quintile, and in both cases raters were off by only one quintile rather than swapping high and low-scoring teachers. Overall, the rank order correlation between observers' ranking and value-added scores was -.47, .02, and .66* in Districts B, G and R, respectively. This is highly variable across districts and, on average, lower than most correlations found in prior studies, suggesting that under the current study's protocol, at least, raters could not accurately predict value-added rank from viewing instruction.

Instructional Features of High or Low Value-Added Teachers

A lack of convergence between qualitative assessments of lessons and value-added scores suggests that there may not be distinct instructional features that characterize high- or low-ranked value-added teachers. However, to explore our second research question, we present data from two sets of analyses. First, we quantitatively describe the relationship between teachers' actual value-added score and whole-scores scores averaged across raters and lessons. Second, we conduct qualitative analyses that draw on initial rater memos and reflection/synthesis of these notes after actual rankings were revealed.

In Table 7, we show the correlations between the items generated for our analysis, averaged across the six lessons scored for each teacher, and teachers' value-added scores. Correlations between value-added and two codes summarizing teachers' overall MQI are of medium strength (r=.37) and statistically significant. Consistent with the literature reviewed above, results from whole-lesson codes that focus on distinct instructional features show that efficiency- and clarity-oriented codes evidence stronger relationships with value-added rankings. Raters' perceptions of whether the teacher used lesson time efficiently had the highest correlation, at 0.45 ($p < 0.05$). The density of the mathematics, the clear launch of tasks, and the clarity of the mathematics all also evidenced moderate and near-significant associations ($p < 0.10$) with value-added scores. One additional item, about whether classroom tasks and activities develop mathematics, had a p-value just outside of significance ($p < 0.11$). In many cases, these correlations were substantively important though not strictly significant, likely due to the small sample size. By contrast, several inquiry-oriented codes, including mathematical inquiry, teacher uses student ideas, and the use of real world situations or examples to motivate student study, did not show either a substantive or significant relationship to teachers' value-added scores.

Qualitative assessments also suggest that, where value-added and observational measures converged, they tended to be around features related to the classroom organization and lesson structure rather than inquiry-oriented instruction. In two districts, this became evident during the ranking phase. In District R, observers' notes indicate that they were fairly certain of value-added rankings for the teachers they correctly assigned to the bottom quintile of value-added scores. Lessons conducted by these teachers suffered from several issues, including mathematical errors and imprecisions, student/classroom disruptions, and in some cases, a lack of focus on mathematics (e.g., reading a mathematically-themed book or playing a mathematically-themed game, but done in a way that generated little mathematical discussion or activity). Two of the classrooms also featured a negative classroom climate, with teachers frequently reprimanding students and students displaying off-task behavior; both teachers appeared frustrated and made statements that raters considered disrespectful to students. Though raters were less certain of the teachers they ranked as high, lacking evidence of inquiry-oriented and meaning-focused lessons, they made decisions regarding classroom organization features as well. Two teachers covered a tremendous amount of material during each class session; students worked continuously on mathematics with very efficient transitions and a very high number of mathematics problems solved (high density). In District G, only two teachers stood out as being clearly situated in a value-added quintile – one in high and one in low. Raters note that the top-ranked teacher used time efficiently, had relatively dense lessons, and built toward a clear mathematical point. Conversely, the low-ranked teacher taught lessons that often were unsystematic, combining unrelated activities with no coherent focus. His lessons also featured more behavioral issues than others in the sample, though not enough to prevent students from appearing comfortable contributing ideas and participating. At the same time, these relationships were true only for a subset of teachers within and across districts. However, classroom organization and management features did not appear to differentiate teachers from District B and were not used to rank them.

In follow-up analyses after actual rankings were disclosed, raters explored the extent to which these features – or others – showed up in high or low value-added teachers instruction. Raters checked several hypotheses generated through discussions about instruction across districts. These included the presence of review material at the start of lessons, the number sense/meaning focus of the lesson, and the presence of a clear mathematical point or purpose. Part of this work focused on uncovering additional common features among teachers whose ranking was predicted correctly, as well as on providing possible explanations for incorrect predictions that may have been due to a focus on too narrow a subset of instructional features (particularly in District B). However, no additional trends were revealed.

Possible Explanations for Misalignment

In addition to describing reasons why some teachers were ranked as high or low, raters discussed several factors that may have contributed to the high degree of misalignment between their predicted and actual guesses.

First, it often was difficult for raters to differentiate teachers when there was little variability in instructional quality. This was true in Districts B and G for all but the two or three teachers predicted to have the highest or lowest value-added score, and in District R for teachers predicted to be mid or high. In District B, two of seven incorrect predictions were cases in which raters noted middling instructional quality and, therefore, were fairly uncertain about ranking in one of two adjacent categories. Similarly, in District G, instruction was fairly uniform with an absence of large variations in clarity and accuracy, apparent teacher content knowledge, classroom climate and behavior management, or pedagogical style. Here, final decisions fell back on small differences in teachers' instruction around the level of student engagement and the cognitive demand of tasks, even though these both were rare.

A related explanation is that, in many instances, each teacher possessed both a range of instructional features and strengths and weaknesses in their implementation, rendering raters unable to prioritize among these when attempting to translate instructional quality into a value-added ranking. In District B, for example, one teacher's instruction included a strong classroom climate in which students were encouraged to take risks (i.e., explaining their thinking and presenting work publicly) and help each other complete tasks; it also contained imprecise teaching of content, particularly around definitions of math terms. In this case, one rater focused on the former feature and ranked the teacher in the highest quintile, while another focused on the latter feature and ranked the teacher in the lowest quintile. This type of dilemma also was true for a number of other teachers. For seven of the nine teachers, at least two of the raters disagreed in their initial value-added predictions. For three of these teachers, raters' predictions spanned all three categories of "high", "mid", and "low". Ultimately, raters sought to reach agreement by focusing on the degree and quality of inquiry-based instruction; even so, they did not agree completely on the final rankings for two teachers.

A third possible explanation regarding misalignment is related to the information available to observers. In District G, many classrooms featured instructional "centers", where students worked in groups on different activities, one of which was with the teacher. Because raters watched videotaped lessons rather than observing them live, the centers format made it difficult to tell how much time students were spending on mathematics, or to identify the content and cognitive demand of activities. As lessons often began with students already in their centers, raters saw very little evidence of how, if at all, these teachers presented new content to their students. Thus, rankings were based on incomplete impressions of what was going on in these classrooms.

Finally, in District B, raters felt that unique classroom compositions (made available to them after making final guesses) might have affected value-added scores but not instructional quality.

One instance was a bilingual class in which the teacher's focus on meaning, language development around mathematical content, and strong rapport with students might not translate to test score gains for these low-performing students. The group predicted that this teacher's ranking was Mid (and even considered high), though the actual ranking was low. A second instance was a teacher that raters confidently rated in the highest quintile because of consistent focus on conceptual understanding of the math. The actual ranking of mid might be related to the fact that students were very high achieving (1.29 standard deviations above the mean in prior average math achievement) and might not have been able to show substantial growth on tests.

Despite these plausible explanations, the fact remains that there was a high degree of prediction error in two of the districts, with seven out of 27 teachers swapping categories from high to low, or vice versa. In District B, raters' predictions were far worse than chance alone, and in fact were negatively related to actual VAM rank. This is particularly surprising given the much wider range in instructional quality in District B than in the other two; wider variability might have resulted in more accurate predictions, as raters were not making distinctions based on small differences in instruction.

Conclusion

This analysis presents mixed findings with regard to the correspondence between instruction and value-added scores. In two districts, raters' predictions were either unrelated or negatively related to actual VAM rankings; in a third, raters' predictions were fairly well matched to VAM rank. We have no explanation for these differences across district; if anything, we would have expected to have the best predictive ability in District B, in which the range of instruction was higher than other districts. One possibility is that District R, in which three classrooms were identified as having low instructional quality through both observational and VAM rankings, raters had the "boost" necessary to correctly predict four of the other six teachers. If so, this might suggest that classroom climate and management are important predictors, at least when classrooms are characterized by poor scores on both.

Taken together, results from this study highlight the difficulty of predicting teachers' value-added through observations of instruction. These results are at somewhat at odds with conventional wisdom and prior work indicating that school leaders are able to identify teachers in the tails, i.e., the "best" and "worst" (Jacob & Lefgren, 2008). Whether principals could do better with a similar sample is an open question; certainly, principals have several advantages (knowledge of classroom context, prior and current student test scores) over external observers. However, observers in this study had several advantages over principals, including multiple observers per lesson and content-knowledge expertise. This issue merits further attention.

More consistent with prior studies, both the quantitative and qualitative analysis suggested that, even with a high degree of misalignment between observations and value-added scores, there were some common elements of instruction that predicted student outcomes – namely, classroom climate and management, efficiency, and clarity and density of the mathematics. Inquiry-oriented instruction and more subtle variables such as the quality of teachers' uses of student ideas were not strongly predictive of outcomes. Coming as we do from a field that has for decades prioritized inquiry-oriented instruction, we find these results surprising and worth further investigation. One possibility is that state standardized tests do not well-detect or reward these classroom features, for student inquiry and reasoning is difficult to measure in conventional formats. Another possibility is that the lack of much inquiry-oriented mathematics instruction in our sample prevented us from detecting the effect of such instruction on outcomes.

Regardless of the performance of inquiry-based items, even in the cases where the items developed for this project predicted VAM outcomes, the resulting correlations were not high. Although our sample size is too small to formally model our variables jointly and to thus calculate an r-squared, we suspect that, given the intercorrelations between these predictor items, the amount of variability in VAM rankings explained did not exceed 0.30-0.40, the range typical in the existing literature. Our qualitative analysis suggests that there may not be one particular instructional dimension on which teachers could be arrayed from strongest to weakest, with that array matching VAM scores even roughly.

These results may suggest that this effort has not shed much additional light on the "production function" that converts classroom teaching into value-added scores. However, given that we have replicated findings across three districts and also given the fact that our results mirror those from the literature, we argue that it may be time to reconceptualize the search for the production function in U.S. classrooms. Rather than seeking to identify specific teacher characteristics or teaching practices that explain a large amount of the variation in student outcomes, we may instead focus attention on how teachers' knowledge and skill interact with one another – and fit the needs of the students in the classroom – to produce valued outcomes. This idea is not new (Cohen, 2012; Cohen, Raudenbush, & Ball, 2003; Hiebert et al., 2005; Lampert, 2001), yet it presents significant challenges to those who wish to create metrics for measuring teachers.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Boonen, T., Van Damme, J., & Onghena, P. (2013). Teacher effects on student achievement in first grade: which aspects matter most? *School Effectiveness and School Improvement*, (ahead-of-print), 1-27.

Brophy, J. (1999). Teaching. *Education practices series*, *1.* International Bureau of Education. Retrieved from http://www.ibe.unesco.org

Brophy, J., & Good, T. L.(1986). Teacher behavior and student achievement. *Handbook of research on teaching*, 328-375.

Chester, M. D. (2003). Multiple Measures and High-Stakes Decisions: A Framework for Combining Measures. *Educational Measurement: Issues and Practice*, *22*(2), 32-41.

Cohen, D.K. (2011). Teaching and its predicaments. Cambridge, MA: Harvard University Press.

Cohen, D.K., Raudenbush, S., & Ball, D.L., (2003). Resources, instruction, and research. *Educational Evaluation and Policy Analysis*, *25*(2), 119–142.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis*, *2*(1), 7-25.

Dee, T., & Wyckoff, J. (2013). *Incentives, Selection, and Teacher Performance: Evidence from IMPACT* (No. w19529). National Bureau of Economic Research.

Fenstermacher, G., & Richardson, V. (2005). On making determinations of quality in teaching. *The Teachers College Record*, *107*(1), 186-213.

Goldhaber, D., & Hansen, M. (2012). Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance. *Economica*.

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (No. w16015). National Bureau of Economic Research.

Grossman, P., & Loeb, S. C. J., & Wyckoff, J.(forthcoming). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added. *American Journal of Education*.

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., ... & Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. *Educational Evaluation and Policy Analysis*, *27*(2), 111-132

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When Rater Reliability Is Not Enough Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher*, *41*(2), 56-64.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Jacob, B., & Lefgren, L. (2006). When principals rate teachers. *Education Next*, *6*(2), 59-69.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101-136.

Kane, T. J., Staiger, D. O., Grissmer, D., & Ladd, H. F. (2002). Volatility in school test scores: Implications for test-based accountability systems. *Brookings papers on education policy*, (5), 235-283

Kane, T. & Staiger, D. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Research Paper, Bill and Melinda Gates Foundation. Retrieved from www.metproject.org

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. National Center on Performance Incentives, Vanderbilt, Peabody College.

Koedel, C., & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. *Education*, *5*(1), 54-81.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education*, *6*(1), 18-42.

Koedel, C., Leatherman, R., & Parsons, E. (2012). Test Measurement Error and Inference from Value-Added Models. *Unpublished Draft*.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven: Yale University Press.

Martinez, J.F. (2013). Combinacion de mediciones de la practica y el desempeño docente: consideraciones technicas y conceptuales para la evaluacion docente. *Pensamiento Revista de Investigación Educacional Latinoamericana*, 2013, 50(1), 4-20.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education*, *4*(4), 572-606.

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-53.

National Council for Teachers of Mathematics (2000). Principles and Standards for School Mathematics. Reston, VA: Author.

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis*, *30*(2), 111-140.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices*. Corwin-volume discounts.

Schacter, J., & Thum, Y. M. (2004). Paying for high-and low-quality teaching. *Economics of Education Review*, *23*(4), 411-430.

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, *38*(2), 142-171.

Stodolsky, S. S. (1988). *The subject matters: Classroom activity in math and social studies*. University of Chicago Press.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*(4), 339-355.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *The American Economic Review*, *100*(2), 256-260.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research, 73*(1), 89-122.

Figures and Tables

Table 1. Sample comparison of students across districts.

	B			G			R		
Demographic	Overall	Study	Sample	Overall	Study	Sample	Overall	Study	Sample
Number of Students	17914	1952	484	76835	3746	643	5384	2252	752
Male	0.49	0.48	0.48	0.50	0.51	0.54	0.51	0.52	0.50
Non-white	0.87	0.93	0.88	0.70	0.70	0.68	0.71	0.72	0.70
Free- or Reduced-Price Lunch Eligible	0.80	0.82	0.81	0.55	0.53	0.49	0.73	0.73	0.73
Special Education Status	0.15	0.15	0.17	0.09	0.09	0.09	0.12	0.12	0.15
Limited English Proficiency	0.33	0.35	0.28	0.18	0.17	0.16	0.23	0.23	0.19
Prior Math Achievement	0.12	0.02	0.17	0.03	0.13	0.40	0.05	-0.03	0.04

Table 2. Items generated during exploratory analysis and used to score the analytic sample

Item	Description
Teacher Uses Student Ideas	Teacher uses student ideas and solutions to move the lesson forward
Teacher Remediates Student Difficulty	Teacher attends to student difficulty with the material
Students are Engaged	Classroom environment is characterized by engagement
Classroom Characterized by Math Inquiry	Students participate in the mathematics of the lesson in a substantive way
Lesson Time Used Efficiently	Lesson time is used efficiently; class is on task, and behavioral issues do not disrupt the flow of the class
Density of the Mathematics is High	"Density" of mathematics is high, in the sense that the class is working through many problems/tasks/concepts and the pace is reasonable or high
Launch of Task	Launch of the mathematical task(s) was mathematically sensible, well-designed, clear and not confusing to children
Mathematics is Clear and Not Distorted	Mathematics of the lesson is clear and not distorted
Tasks and Activities Develop Math	The tasks and activities done by the class contribute to the development of mathematical ideas, procedures, etc.
Students Conduct Error Analysis	Teacher asks students to conduct error analysis
Teacher Uses Real World Examples	Teacher motivates mathematical definitions/procedures by using real-world examples
Whole-Lesson MQI	Holistic score of the quality of the mathematics of the lesson, including the richness of the teacher's and students' mathematical productions, the productivity of teacher and student mathematical interactions, and the precision of the teacher's mathematics

Table 3. Holistic code summary statistics

	Overall, N=27		District B, N=9		District G, N=9		District R, N=9	
Holistic Code	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Teacher Uses Student Ideas	2.73	0.60	3.12	0.59	2.74	0.55	2.33	0.41
Teacher Remediates Student Difficulty	3.30	0.43	3.52	0.33	3.33	0.43	3.06	0.44
Students are Engaged	3.28	0.45	3.29	0.58	3.35	0.37	3.20	0.41
Classroom Characterized by Math Inquiry	2.74	0.74	3.29	0.70	2.75	0.55	2.17	0.52
Lesson Time Used Efficiently	2.86	0.53	2.91	0.49	2.81	0.50	2.85	0.65
Density of the Mathematics is High	2.83	0.56	2.85	0.52	2.72	0.39	2.93	0.74
Launch of Task	2.96	0.27	2.99	0.25	3.01	0.17	2.88	0.36
Mathematics is Clear and Not Distorted	3.26	0.59	3.60	0.45	3.18	0.34	3.00	0.78
Tasks and Activities Develop Math	2.91	0.51	3.13	0.38	2.81	0.33	2.80	0.71
Students Conduct Error Analysis	0.03	0.06	0.04	0.07	0.02	0.03	0.02	0.07
Teacher Uses Real World Examples	0.26	0.19	0.22	0.20	0.33	0.20	0.22	0.18
Whole-Lesson MQI	2.89	0.50	3.02	0.55	2.90	0.30	2.76	0.62

Table 4. Within-1 agreement rates for holistic codes

		District		
Holistic Code	Overall	B	G	R
Teacher Uses Student Ideas	0.81	0.83	0.83	0.74
Teacher Remediates Student Difficulty	0.90	0.85	0.95	0.89
Students are Engaged	0.89	0.87	0.95	0.82
Classroom Characterized by Math Inquiry	0.84	0.86	0.84	0.81
Lesson Time Used Efficiently	0.91	0.88	0.92	0.93
Density of the Mathematics is High	0.86	0.77	0.94	0.86
Launch of Task	0.93	0.88	0.98	0.92
Mathematics is Clear and Not Distorted	0.83	0.81	0.87	0.79
Tasks and Activities Develop Math	0.91	0.94	0.91	0.88
Whole-Lesson MQI	0.90	0.87	0.91	0.92

Table 5. Variance decomposition of holistic codes

Holistic Code	*d*	*t: d*	*E*
Teacher Uses Student Ideas	0.09	0.24	0.67
Teacher Remediates Student Difficulty	0.05	0.37	0.58
Students are Engaged	0.00	0.41	0.59
Classroom Characterized by Math Inquiry	0.20	0.35	0.45
Lesson Time Used Efficiently	0.00	0.44	0.56
Density of the Mathematics is High	0.00	0.39	0.61
Launch of Task	0.00	0.20	0.80
Mathematics is Clear and Not Distorted	0.04	0.35	0.61
Tasks and Activities Develop Math	0.00	0.38	0.62
Students Conduct Error Analysis	0.00	0.07	0.93
Teacher Uses Real World Examples	0.00	0.12	0.88
Whole-Lesson MQI	0.00	0.43	0.57

Note. The facets of variance include the district (*d*) level, the teacher nested within district level (*t : d)*, and the residual, or lesson nested within teacher nested within district, level (*e*).

Table 6. *Percentage of incorrect teacher VA quintile guesses*

District	Percent	Guesses
Off By 1+		
B	77%	9
G	44%	9
R	22%	9
Overall	48%	27
Off By 2		
B	50%	6
G	66%	6
R	0%	6
Overall	39%	18

Table 7. Correlation coefficients between teacher scores on study-generated items and value-added scores

Quantitative Code	Value-Added
Teacher Uses Student Ideas	0.01
Teacher Remediates Student Difficulty	0.26
Students are Engaged	0.12
Classroom Characterized by Math Inquiry	-0.08
Lesson Time Used Efficiently	0.45*
Density of the Mathematics is High	0.35~
Launch of Task	0.35~
Mathematics is Clear and Not Distorted	0.34~
Tasks and Activities Develop Math	0.31
Students Conduct Error Analysis	0.18
Teacher Uses Real World Examples	-0.22
Whole-Lesson MQI	0.37~

~ $p < .10$

* $p < .05$

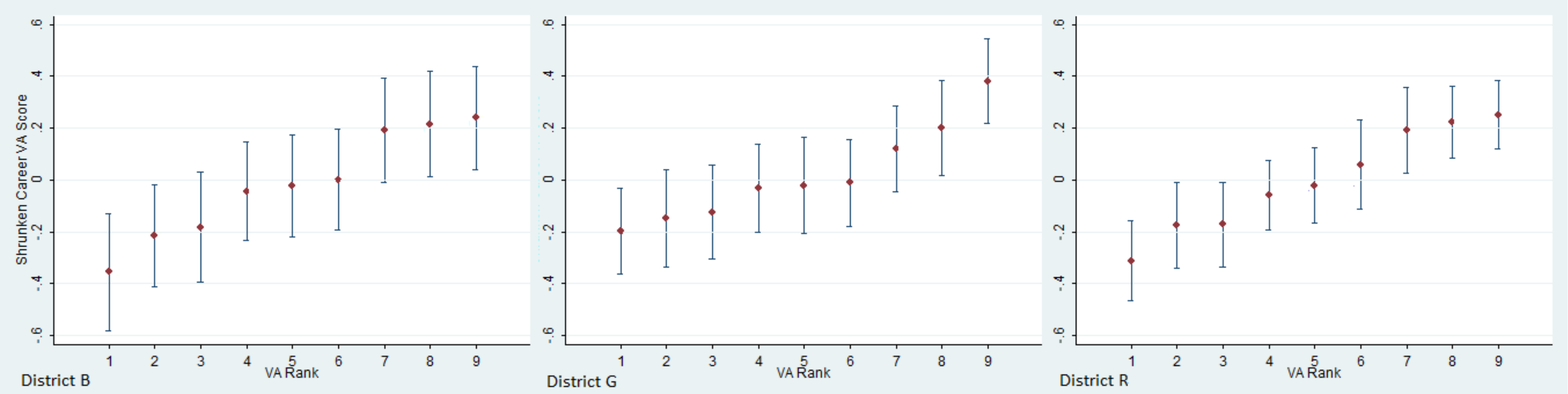

Figure 1. Range plot with capped spikes showing confidence intervals of value-added scores for teachers in each district

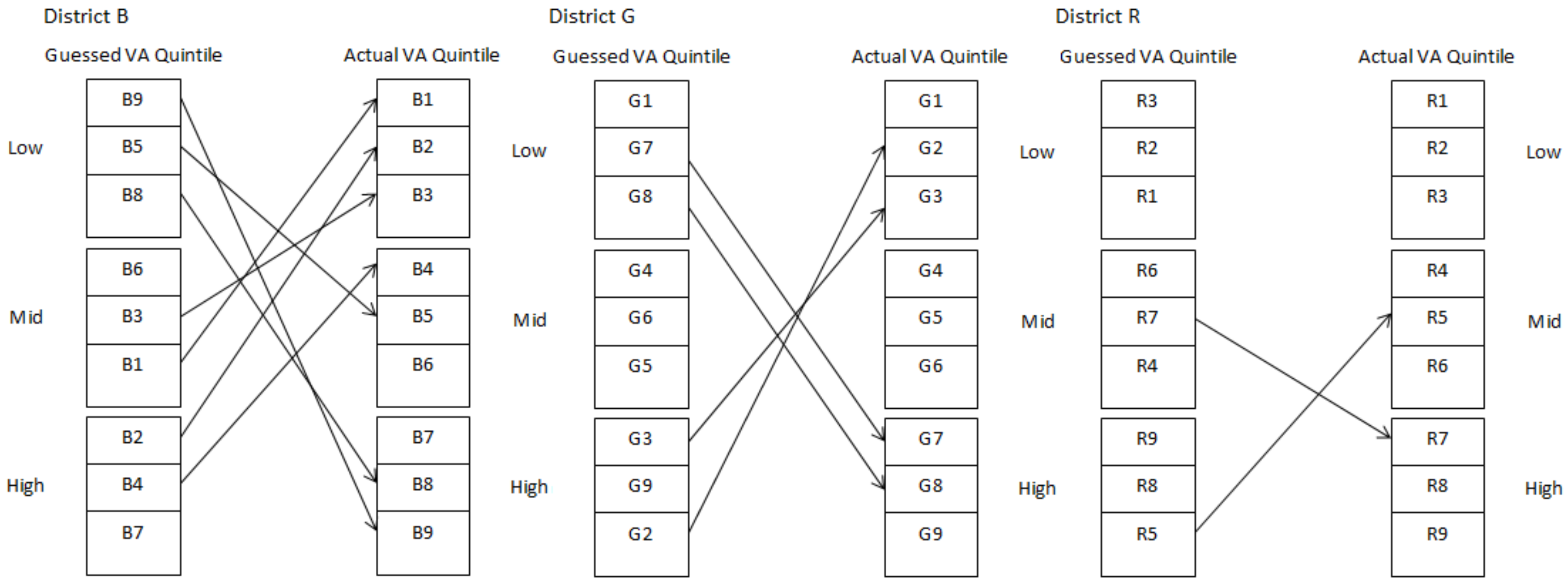

Figure 2. Comparison of teacher VA quintile guesses to actual VA quintile, for all districts.